ICAME Journal • 47(1) • 2023 • 119–139
DOI: 10.2478/icame-2023-0007

sciendo

ICAME Journal

Compiling a corpus of South Asian online Englishes: A report, some reflections and a pilot study

Research Article

Muhammad Shakir*, Dagmar Deuber

University of Münster

Received 8 November 2022; accepted 2 January 2023

Abstract: In this research article we introduce the **S**outh **A**sian **On**line **E**nglishes (SAOnE) corpus representing four South Asian countries, i.e. Bangladesh, India, Pakistan, and Sri Lanka, and two native English-speaking countries, i.e. the UK and the USA. We have used semi-automatic and manual methods to collect data from three internet registers, i.e. newspaper comments, web forums and tweets, and a collection of internet sub-registers which we label as blogs and websites. Additionally, we have collected text messages using online freelance hiring platforms from each of the South Asian countries mentioned above. Each register category in the corpus consists of approximately 1 million words per register per country, except text messages, which contains around 500,000 words per country and only includes the four South Asian countries. We have verified the origin of website and blog links, authors of Twitter, and where possible of commenters and web forum users to make sure that only local content of each country is included. The corpus features some indigenous language content, which is tagged.

In addition to the description of this dataset, we also present a pilot study analysing three discourse particles, namely *na*, *neh*, and *yaar*. The discourse particles *na* and *yaar* are native to Hindi/Urdu, while *neh* is based on a Sinhala negation marker. Our analysis indicates that *na* and *neh* have similarities in terms of their position in the clause/utterance. However, *neh* is confined to Sri Lanka while the Hindi/Urdu based discourse particles are also used in our Twitter data from Sri Lanka and Bangladesh. The use of these discourse particles in Bangladeshi tweets shows the influence of Indian culture through Bollywood celebrities. Of the Hindi/Urdu discourse particles *yaar* and *na*, *yaar* is preferred in Pakistan while *na* is preferred in India; additionally, *yaar* is used at the start of the clause more often in our Pakistani data. Lastly, we discuss the implications of the pilot study, the advantages of the type of data used for the pilot study, and future research directions.

Keywords: *SAOnE corpus • internet register • newspaper • web forum • tweet • blog • website • discourse particle*

1 Introduction

Corpora are one of the most useful tools for the study of varieties of English around the world. Though corpora of traditional spoken and written registers have been available for numerous varieties of English for a long time, there are not many datasets consisting of a diverse set of internet-based registers, especially for non-native varieties of English like those in the South Asian region, namely Bangladesh, India, Pakistan, and Sri Lanka. This is a notable gap not only because internet-based registers are an important part of today's communicative landscape but also because they potentially include informal and interactive communication. The latter is hard to find in the form of traditional spoken-informal registers like face-to-face conversations in countries like Bangladesh, where the use of English is quite restricted (Chowdhury and Kabir 2014). Moreover, being multilingual, South Asian users often code-switch in their English communication, so interactive online registers also present an opportunity to study this phenomenon. The presently available datasets do not include the type of internet-based registers

* *Corresponding author: Muhammad Shakir, E-mail: muhammad.shakir@uni-muenster.de*

Open Access. © 2023 Muhammad Shakir, Dagmar Deuber, published by Sciendo.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

discussed above, e.g. text messages. We propose the compilation of such a corpus with at least a few million words per country, which will be several times bigger than the individual ICE components, e.g. ICE-India (Shastri and Leitner 2002), but much smaller as compared to the very large corpora currently available, e.g. GloWbE India (Davies and Fuchs 2015), which has 96,430,888 words. The in-between size category of this corpus will permit the application of ICE-like principles in corpus compilation, and hence make the data comparatively more reliable. Such a corpus should be well suited to throwing more light on the development of English on the internet. Especially for South Asian varieties of English it holds major potential for research, as we are going to illustrate in this paper.

Our first aim is to present the South Asian Online Englishes (SAOnE) corpus that has been compiled for the above-mentioned South Asian countries along with two reference countries, i.e. the United States and the United Kingdom, containing the hyper-central and the historically relevant super-central variety of English (Mair 2013: 261), respectively. SAOnE consists of openly available web registers, i.e. general websites and blogs, web forums, newspaper comments and social media data from Twitter, as well as text messages from various platforms like WhatsApp, Telegram, and Facebook Messenger.[1] Our second aim is to demonstrate the usefulness of this dataset by presenting a pilot study of three indigenous discourse particles, *yaar*, *na*, and *neh,* that are employed in the English communication of South Asian internet users.

The rest of this paper is structured as follows. In the next section, we provide an overview of the current state of affairs regarding internet-based corpora and the rationale behind the development of this corpus. In Section 3 details of data collection and corpus compilation are described. Section 4 presents the pilot study of the three discourse particles mentioned above. We conclude by discussing implications, limitations, and future research directions.

2 Background

An important factor in research on varieties of English is the availability of a sizeable corpus. In the South Asian context, the oldest corpora include the Kolhapur Corpus (Leitner 1991) of Indian English and ICE (International Corpus of English) India (Shastri and Leitner 2002); ICE Sri Lanka (Mukherjee et al. 2010) is under construction and the status of ICE Pakistan is unknown (Kirk and Nelson 2018; an unofficial version is, however, available), while an ICE component for Bangladesh does not appear feasible due to the very restricted use of English in this country. Consequently, Indian English is one of the most researched varieties within South Asia. There are of course also other reasons for this apart from corpus availability, e.g. the size of the country, the influence of Indian English on other South Asian varieties as well as the presence of many sub-varieties within Indian English. The study of other South Asian varieties of English has been facilitated by the availability of the South Asian Varieties of English (SAVE) corpus (Mukherjee and Schilk 2012), but this dataset lacks the diversity of the ICE as it only consists of online newspaper data.

With the popularity of the internet in the last two decades, linguistic researchers and corpus compilers have increasingly turned their attention to internet-specific corpora. One of the most widely used datasets in World Englishes research is the Global Web-based Englishes (GloWbE) corpus (Davies and Fuchs 2015). GloWbE contains data from blogs and other websites for 20 varieties of English, including the South Asian varieties. Despite its immense usefulness, the mega corpus also has its shortcomings: the assignment of texts to countries lacks reliability as a result of reliance on the search engine Google and domain information (Nelson 2015). The inclusion of “unintelligible and uninterpretable gibberish” (Mair 2015: 31) is another issue. Additionally, despite its enormous size, the corpus lacks the systematicity of internal variation that is readily available in small carefully compiled datasets like the ICE (cf. Loureiro-Porto 2017 for a comparison of ICE and GloWbE). For example, our own observations indicate the

1 SAOnE can be pronounced as /sa:vən/, inspired by the word Sawan in South Asian languages like Punjabi and Hindi/Urdu (it refers to the fifth month of the Vikrami Calendar).

presence of web forums both in the Websites section as well as in the Blogs section of GloWbE, which makes it difficult for a researcher in case they are looking to target only informal/ interactive registers (cf. 4.2). There have also been individual attempts to collect social media data for researching varieties of English. For example, Bohmann (2019) collected tweets from various countries for his study. Such datasets, however, are not diverse in terms of register selection and do no employ an extensive procedure to obtain a more reliable dataset as we explain below for our corpus.

The idea of an ICE-style online corpus of varieties of English is not new. Kirk and Nelson (2018: 716) have proposed ICE add-on corpora of the following internet-based registers: Websites, Email, Tweets, Chat, Discussion groups, Facebook posts, Skype, Blogs, Vlogs and podcasts, and Citizen broadcasting. While the authors do propose a diverse set of internet-based registers, there are a number of issues with their proposed corpus structure. Text categories related to Text/SMS and Chat have either been merged in today's instant messaging apps like WhatsApp, or they do not exist in the original form, i.e. the 140-character text messages of the late 2000s. Certain categories like Vlogs and podcasts etc. are difficult to find for South Asian countries like Bangladesh or Pakistan, where such publications will most probably occur, if at all, in local languages because the majority of the population does not understand English and the content creators need viewership to earn advertisement revenue. Social media related categories like Facebook postings also present the challenge of copyright compliance: Facebook does not allow collection or redistribution of its data. Additionally, the user privacy and data protections rights are also a considerable hurdle. Lastly, the proposal overlooks an interactive online register that is among the oldest ones and still commonly available, namely comments (e.g. from newspaper websites or social media websites like YouTube).

Based on the above discussion, the present corpus should include a diverse range of registers, especially from the informal/interactive end of computer-mediated communication. Additionally, the country of origin of the data should be reasonably verified to increase reliability. Although the internet presents a wide, ever-expanding and ever-changing range of communication possibilities, the choices of the selection of registers are limited by a number of factors. The first and foremost is the accessibility – as in public, open and programmatically accessible – of registers for easier data collection. The websites (World Wide Web) that can be opened in a browser and do not require a login are the most obvious choice. Biber et al. (2015) identify eight general and more than 20 sub-registers, including discussion-oriented registers, i.e. discussion forums and comments, on the World Wide Web. The second factor is the verifiability of the country of origin. Discussion forums and a sub-type of comments, i.e. the ones occurring on a newspaper website of a particular country, can be reasonably expected to have a majority of users from the said country. Similarly, manual verification of origin can also be performed for the other sub-registers published in the form of webpages. Hence, these registers are good candidates for inclusion in the corpus.

The most interactive/informal and multimodal registers occur on websites and platforms that are known as social media websites/apps, e.g. Facebook, Instagram, TikTok, Youtube, WhatsApp, Telegram etc. However, these platforms are generally not accessible in the sense as described above. Even though data can be collected from, for example Instagram (cf. Sardinha 2022), such data cannot be included in a re-distributable corpus. The data from Twitter as tweets and Youtube as public comments on videos can be considered accessible, although these platforms only grant access through an application programming interface or API. Out of these two platforms, it is easier to verify the origin of Twitter users as compared to those of Youtube because the latter do not have profile information readily available. Additionally, Twitter data is redistributable with some restrictions (Twitter 2022), which makes this platform more suitable for the present project. Highly synchronous interactive registers are available on messaging platforms like WhatsApp, Facebook Messenger, and Telegram, which are not publicly accessible. However, individual users can be requested to donate their data and their country of origin can also be verified in this process. Platforms like TikTok and Youtube present many new formats of multimodal communication on the internet, e.g. short videos and podcasts. A third factor becomes relevant here, i.e. the availability of a particular register in a South Asian country. Since the availability of registers like podcasts in English cannot be guaranteed for all South Asian countries, such multimodal

registers become less attractive for the present data collection. Hence, in terms of interactive registers from social media platforms, two registers, i.e. tweets and text messages, appear to be suitable for inclusion in the present corpus. The following section provides details of the SAOnE corpus and the corpus compilation process.

3 Corpus compilation

3.1 Design of the corpus

The corpus consists of English communication from four South Asian countries: Indian English, which is the super-central variety in this region, and Pakistani, Sri Lankan and Bangladeshi Englishes, which are central varieties of English (Mair 2013: 261). Additionally, the hyper-central variety, i.e. US English, and the historically relevant super-central variety, i.e. UK English (Mair 2013: 261), have been included as reference to facilitate linguistic comparisons that we plan for these data and for any future researchers. In terms of the text categories, as discussed in the previous section, our decisions have mainly been guided by three main factors: accessibility of data (open, publicly accessible, and programmatically downloadable), verifiability of the country of origin, and high chances of availability in all four South Asian countries. Additionally, we include registers with no or minimum copyright restrictions in case of social media data, hence Twitter. For text messages, we ask for group chats on study-related or professional topics, i.e. semi-private chats, to increase our chances of data collection.

Table 1 shows the corpus categories and a brief description of data sources for each category. Table 2 exhibits that we have aimed for 1 million words per country in most categories, but the data collected depends on availability; for example, not enough news comments were available for Bangladesh. Blogs were merged with general websites mainly because there was not enough data available from Bangladeshi and Sri Lankan blogs.[2] In this category, the first 100 files (20 percent of the whole) – or less in the case of Bangladesh and Sri Lanka – contain blogs data, while the rest are from general websites which may also include blogs. Text messages were limited to 500,000 words due to the additional time and effort required to acquire, clean, and anonymise these data. No text messages were obtained from the USA and UK, as efforts to recruit help for data collection proved futile (see below for more details).

3.2 Data collection

For text messages, we started the data collection process by hiring freelancers from the four South Asian countries using online freelancer hiring platforms, initially Fiverr and later mainly Upwork.com. Advertisements as shown in Figure 1 were posted for each South Asian country and also, unsuccessfully, for the USA and UK. The potential freelance assistants were provided with an introductory document explaining the aims and objectives of the project. A consent form was drafted and the freelance helpers were instructed to share it with the participants (cf. the text in Figure 1). The same consent form was submitted by the freelancers on behalf of the individual participants while uploading the chat histories through a Google form. Initially, several dozen freelancers were hired and small data samples were collected from them. In the second stage, the number of freelancers was reduced to a few individuals per

2 As mentioned previously, Biber et al. (2015) find multiple sub-registers on the web, which have been identified by raters hired through a crowd-sourcing website. In spite of a decision tree like task, they ended up with a considerable portion of hybrid registers, i.e. one text having more than one register labels. Later, Biber and Egbert (2018: 196) propose the idea of "the web as a continuous space of register variation" especially for these hybrid registers, where each text can have multiple communicative purposes and situational characteristics with no clear-cut register boundary. The websites data collected in SAOnE should also have similar tendencies, with only web forums and reader comments readily identifiable. It entails that human raters are needed to check each text and decide the register category or at least its situational characteristics. Since we focus on the verification of the country of origin, checking for copied content as well as elimination of spam content, we do not have enough resources to allocate for the identification of sub-registers in the data collected from websites. Additionally, we do not expect that all sub-registers, as proposed for example in Biber et al. (2015), will be present in all South Asian countries to create balanced sub-categories. Due to these reasons, we categorise all static webpages and blogs etc. under Blogs and Websites, which excludes any interactive registers like reader comments or forum posts. This is an area which we would like to improve on in a future research project, i.e. identifying sub-registers in all or a sample of the texts in Blogs and Websites in each country.

Table 1. Overview of the corpus categories and their description.

Register category	Description
Comments	• Newspaper comments (2 newspapers per country from 4 SA countries + USA and UK). • Comment threads combined in ≈2000-word files.
Discussion forums	• Discussion threads from public English forums, at least 5 web forums per country on topics like cars, technology, cricket, defence etc. from 4 SA countries + USA and UK. • Forum threads combined in ≈2000-word files.
Tweets	• English tweets originating from 4 SA countries + USA and UK as provided by Twitter API. • 1 user/file in ≈2000-word files.
Text messages	• Mostly English group and 1-1 chats on topics like study, work etc. from various messaging platforms (WhatsApp, Facebook Messenger, Telegram, Upwork…) originating from 4 SA countries. • Complete conversations per file, spanning from a few hundred to 100,000 or more words.
Blogs and websites	• English blogs identified from blog directories, general websites in English through links generated from Google (4 SA countries + USA and UK). • Posts/page texts combined in ≈2000-word files.

Table 2. Word count and number of files in each corpus category and country.

Register category	Bangladesh	India	Pakistan	Sri Lanka	UK	USA
Comments	579965 (284)*	1016502 (500)	1012037 (500)	1009063 (500)	1024066 (500)	1027162 (500)
Discussion forums	1035544 (501)	1080612 (504)	1074940 (505)	1034034 (504)	1063475 (504)	1059083 (504)
Tweets	1078517 (500)	1107505 (500)	1079919 (500)	1063949 (500)	1073723 (500)	1053780 (500)
Text messages	575094~ (84)	555200 (43)	548874 (129)	599344 (75)	0	0
Blogs and websites	1019670 (500)	1015792 (500)	1016759 (500)	1023404 (500)	1013988 (500)	1014653 (500)
Total	4288790 (1869)	4775611 (2047)	4732529 (2134)	4729794 (2079)	4175252 (2004)	4154678 (2004)

* Words (Files); ~ Word count includes timestamps for text messages

country who provided quality data. Generally, we aimed to include chats where at least 70–80 percent of the discussion took place in English. We asked for group chats of students or university class groups, office or workplace WhatsApp or Facebook Messenger groups, or one-to-one chats if the freelance assistant could find such text messages. In addition to the chat histories, the freelance helpers were asked general questions related to the background of the participants; see Table 3 for a summary of details.

Web forums and newspaper comments were scraped using the package Scrapy (Kouzis-Loukas 2016) in Python (Python Core Team 2019). We identified five or more web forums from each South Asian country using Google. Any forum post published in 2011 or after this year was scraped. Where applicable, we tried excluding sub-forums where there was a high probability of foreign forum-posters; for example, on the Defence.pk web forum we excluded sub-forums related to India or other countries due to the presence of non-Pakistani users. Also, sub-forums were scraped equally to ensure diversity of topics within the given web forum. The same was repeated for the USA and UK. For these two countries, users' location information was sometimes available, which was duly used to exclude non-local users in the scraping process. For comments, we identified two newspapers from each country and downloaded the comments under discussion-inducing categories like opinion articles, letters to the editor, and blogs, but also news reports when not enough data was available. All comments published in the year 2011 or later were included. The 500 corpus files for each category were equally divided to each source. For example, for Bangladesh web forums, the 500 files were equally divided into eight web forums (approximately 63 files per source). If a source did not have a sufficient number of files, the remaining files were equally divided among the yet unused sources. In the case of comments, 250 files were created from each newspaper unless a newspaper did not have enough comments, in which

Table 3. Summary of meta information about text messages (conversation/ file as the unit).

Information category	Summary of answers
Messaging platform*	WhatsApp (77.4%), Facebook Messenger (13.5%), Telegram and others (9.1%)
Chat type	Group chat (71.6%), one-to-one chat (28.4%)
Participant location (multiple answers possible)	Most frequent: South Asian countries (97%) Least frequent: UK, USA, Australia
Participant relationship (multiple answers possible)**	Most frequent: Friends including class follows (76.6%), colleagues (22%) Least frequent: Family (8.3%), clients (7.2%), strangers (6.3%)
Age groups (multiple answers possible)	Most frequent: 18-25 (67.8%), 26-35 (45.7%), 36-45 (13.2%) Least frequent: 46+ (4.4%), 10-17 (0.3%, 1 instance)
Participant gender	Mixed (65.8%), male only (20.9%), female only (13.2%)
Indigenous languages possibly present in text messages (multiple answers possible)***	Bangla, Hindi, Kannada, Kashmiri, Malayalam, Pashto, Punjabi, Sindhi, Sinhala, Saraiki, Tamil, Telugu, Urdu

*Percentages have been calculated based on number of texts or files (1 file = 1 conversation), e.g. 77 percent WhatsApp means that 77 percent of texts or conversations are from WhatsApp texting platform.
**For example, a file can include 'Friends including class follows' as well as 'strangers'. The file is counted in the percentage value of Friends including class follows (76.6%) as well as in the percentage value of strangers (6.3%). This is the reason percentages do not sum to 100 percent.
***The information is provided by the freelancers while submitting the conversation text files through a Google Form. Sample form: https://docs.google.com/forms/d/e/1FAIpQLSf8wYfQ9h031NUAS5fznE8QJzuGrs_A-nAp2Ha_ZDXtMA5EuQ/viewform?usp=sf_link.

case more than 250 files were included from the other source (e.g. in the case of Sri Lanka). Since the comments and web forum posts could also be in an indigenous language, especially in the case of Bangladesh, the langdetect Python module (Danilak 2021) was used to detect the language of the respective comment or forum post. Such comments or forum posts in the indigenous language were allowed in the text files, but they were not included in the final word count.

Blogs data were compiled in two steps. In the first step, the links were generated. Blogs were found through blog directories like Blogflux (http://dir.blogflux.com/), where location-based blog listings were available. Additionally, search queries like "best bloggers from Bangladesh" were used on Google to get any additional blog lists compiled by local bloggers especially for smaller countries like Bangladesh and Sri Lanka for which not enough blogs could be found through the blog directories.[3] Due to the discontinuation of Google Blog Search, this tool could not be utilised. Freelance assistants were then asked to get at least ten blog post links per selected blog from each country in the case of a single-writer blog. More than ten blog post links were obtained from multi-writer blogs (e.g. Pakpositive.com) and news blogs (e.g. Times of India blogs). All available links in the case of multi-writer blogs from Sri Lanka were harvested because not enough blogs were available. In the second step, the blog posts were downloaded using Python and various web browsers. The downloaded data were used to generate blog text files by combining blog posts from the same author and trimming the last added blog post to make the cut at 2000 words. If not enough text was available, multiple blogs or blog authors were combined in the same file. In total, 100 text files were created for the USA, UK, India, and Pakistan, while 60 text files could be created for Sri Lanka and only 34 for Bangladesh.

The data from websites were also collected in two steps. As the first step, freelance assistants were asked to set the respective country as their location in Google Advanced Search and use the most frequent trigrams of function words generated from the Corpus of Contemporary American English (cf. Davies and Fuchs 2015: 4) as search queries. 8000 or more links were generated from each country. Just like blogs, Python and web browsers were used to download the text from these HTML documents. In the second step, 2000-word long text files were generated. Where possible, web documents belonging to the same top-level domain were put in one text file.[4] The last added web document was trimmed to

3 Using such a search query is prone to bias, but it was the best way we could think of to access local blogs.
4 Of course, same top-level domain does not mean the text file contains same type of sub-registers. It was mainly to put similar links next to each other and facilitate the freelance assistants in their review work.

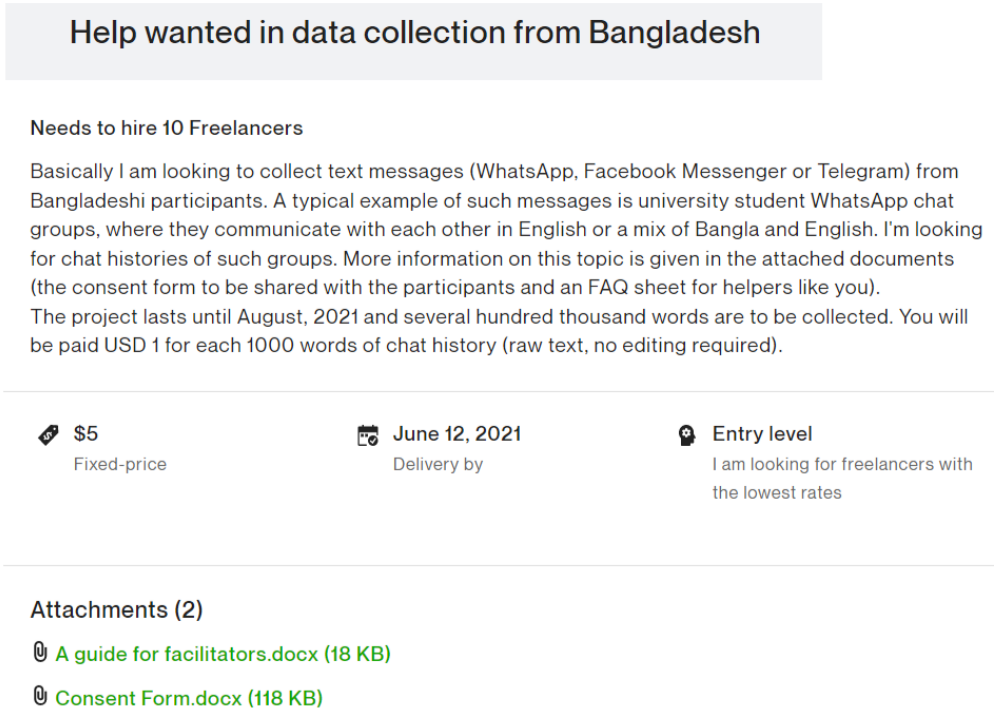


Figure 1. A job post on upwork.com to hire freelancers for text messages data collection from Bangladesh.

match the 2000-word limit. Links were obtained for the webpages that were included in the text files. Our freelance assistants visited the webpages using these links. Any nonlocal webpages, webpages from international websites, spam webpages, text from PDF or other non-HTML files, and interactive websites like web forums and/or question and answer websites were excluded based on this check. The discarded webpages were replaced with newer ones which were also verified for the same.

The text files for Twitter data were generated based on individual users (i.e. one user per text file), so the first step in this regard was to identify users from each country to be included in the corpus. We used Twitter API's live stream function to collect English tweets from each country. Approximately, one thousand users were identified from each country, except for Bangladesh, for which we had to generate additional users in later stages, so the overall number of users identified for Bangladesh increased to around 1800 (see Section 3.3, paragraph 2). The users obtained in this way were manually checked by freelance assistants by looking at their profile information, location, names etc. to make sure that they were native residents of the respective country. Around 100 users were removed as a result of manual screening and we were left with about 800 users per country. Afterwards, approximately 350 English tweets were downloaded using the Twitter API for each user in the final list. After downloading the tweets, we randomly selected 500 users from each country to create 2000-word long text files consisting of approximately 130 tweets per user (cf. Table 2 for corpus composition).

3.3 Corpus annotation

After the creation of corpus files in each category for each country, the next step was to manually annotate the corpus for the phenomena detailed in Table 4. We highlight two main issues here. One of the issues was related to the tagging of indigenous content. Local freelance assistants from each country were selected to go through the data and annotate parts that were indigenous. The files were then reviewed by the first author, a student assistant or a freelance assistant to make sure the tags were correctly applied. The other noteworthy and perhaps the most labour and time intensive task was

Table 4. Annotation scheme and usage description.

Tag	Use cases
$A, $B…$Z, $A1, $B1…, $Z1, $A2, $B2… $S3P1, $S3P2, $S3P3… +	Nicknames for text message participants Nicknames for third persons mentioned in the chat
<anon>*	In place of people's names, phone numbers, any personally identifiable information in text messages.
<robotic>	Automatic messages added by WhatsApp, Telegram bots, web forum software when a post is edited, button labels on websites, attachment (file) names in text messages and web forums
<unclear></unclear>	Very rarely used when it is not clear if a word is indigenous/foreign or an English word.
<unclear type="proper noun"> </unclear>	When a participant name is not attributable to their phone number, or vice versa (after anonymising the phone number to a few digits) in text messages.**
<repeat>	If a whole tweet, forum post, or comment occurs a second or third time, i.e. identical content, this tag is placed after removing the repeated content.***
<indig></indig>	Single words (greetings, discourse particles e.g. *yaar*, local festivals, kinship terminology, month names like Islamic month names…), multiword elements (long proper nouns, forum posts, tweets, forwarded local content in Pakistani text messages etc.)
<foreign></foreign>	Non-English and non-indigenous content, for example German in Sri Lankan text messages, Hindi in Sri Lankan Tweets and forums, Spanish in US comments.
<quote></quote>	English copied content from local website of respective South Asian country in text messages, forums etc., text messages composed elsewhere (official notifications, parts of student assignments etc.)
<quote type="foreign website"/>	In place of the content copied from a foreign website in text messages and web forums, motivational and celebrity quotes in tweets, text copied from foreign websites or non-local websites
<foreign type="poster origin"/>	In place of the comment or forum post whose author/commenter does not belong to the particular country, especially in the case of Bangladeshi, Indian and Pakistani news websites.****
<quote type="potentially spam"/>	In place of advertisement related comments especially in the case of Bangladesh, web texts consisting of long lists of keywords or key phrases for search engine optimisation (SEO) making no sense whatsoever
<quote type="code"/>	In place of computer code or HTML code on web forums, text messages, websites (failed to convert to text)

*Such tags will be replaced with self-closing tags in the final version of the corpus.
**Nicknames are assigned based on the phone number or name saved in the data donor's phone, which appear in the timestamp before the message. When a name or a phone number does not appear in the timestamp (i.e. the said person did not send a single text in the whole conversation) but as per the context they are part of the group or conversation, it is very hard to guess who is who. In such cases we apply this tag on the names and phone numbers. In the example below, the people mentioned with names are clearly part of the conversation but either they did not appear anywhere in timestamps or their phone numbers were visible instead of names, so we could not match the names with phone numbers. Hence, replacing, for example, Muneeb with a nickname was not possible:
03/12/2020, 12:47 pm - $F: <unclear type="proper noun">Muneeb</unclear> and <unclear type="proper noun">Fatima</unclear>, plz join the meeting now.
In this example the phone numbers are given and not the names. Class representatives (CRs) are always part of the group, but since we cannot identify who is who, we just tag these phone numbers after anonymising them.
02/12/2020, 9:54 pm - $T1: Who's CR?
02/12/2020, 9:56 pm - $U1: <unclear type="proper noun">@92*****</unclear>and <unclear type="proper noun">@92*******</unclear>
***It does not indicate a quotation as possible, for example, in discussion forums when a user replies to a previous post. Such quotations were programmatically excluded at the downloading stage to remove identical content.
***Such commenters were mostly identified using their nicknames. For example, "Eric Blanco" in India or Bangladesh, or "Abhinav" in the case of Pakistan are presumed foreign commenters. Sometimes they gave an indication like "from India" in their comment. Other times they positioned themselves in a way which showed they do not belong to the respective country, e.g. "I have visited Chittagong 3 times…" with a foreign nickname "Peter Ingram".

to annotate copied content that existed in various forms, i.e. content posted by foreign commenters; forwarded content in text messages; copied articles and news reports in web forums; content from foreign websites, e.g. news reports from foreign news agencies, published on local news websites or blogs; inspirational and celebrity quotations, robotic tweets, and content from news articles on Twitter etc.

The issue of copied content led to the rejection of a substantial amount of text. To compensate for the lost content, the files were revised at least two times in all corpus categories (blogs and websites, comments, web forums, and Twitter) of all South Asian countries and at least once for the USA and UK

(comments, forums, and Twitter). The copied content was tagged by freelance assistants in the first step (e.g. <quote type="foreign website"></quote>), and in the second step it was programmatically removed and replaced with a self-closing tag (e.g. <quote type="foreign website"/>).[5] New blog posts or web texts, comments, forum posts, or tweets from the same website (where possible same blog author, comment thread or forum thread) or the same Twitter user were added to the file. Sometimes, especially in the case of Bangladesh, a Twitter user was simply replaced with another user from the same country if the number of English tweets became too low to make up 2000 words. The newly added content was again edited by freelance assistants as well as reviewed. The copied content identified in the newly added parts was again replaced with yet another batch of fresh content, and the cycle started again. In the last iteration, a file was kept if it contained at least 1000 words and the process of 'file repairing' was ended in the interest of time.[6]

3.4 Indigenous and discarded content in the corpus

As mentioned above, one of the basic ideas behind our corpus design was to also include non-English, i.e. indigenous, content in the corpus. Such content occurs either mixed with English content, for example in text messages, or sometimes the entire posts, comments, or tweets had been tagged as indigenous during the annotation process. Figure 2 shows the percentage of words tagged <indig></indig> in various corpus categories in each of the four South Asian countries. As can be seen, forums, tweets and text messages, all contain some indigenous content, with text messages having the most, and tweets and forum posts less. The highest percentage of indigenous content is found in Pakistani text messages, which is mainly due to forwarded messages copied from social media (mainly Facebook) posts, local Urdu news websites, and in some cases content from books or magazines. The indigenous content in the Bangladeshi, Indian, and Sri Lankan text messages is attributable to code-switching by the chat participants, which occurs in the Pakistani text messages as well. The Pakistani and Bangladeshi web forums sometimes contained posts in indigenous languages (i.e. Urdu or Bangla respectively), which resulted in about 5% of indigenous content in both of these countries. Lastly, the higher percentage of indigenous content in the case of tweets in India and Pakistan is partly because we relied on Twitter API's language detection algorithm to get English tweets. Many tweets in Hindi or Urdu written in Roman script were misclassified by Twitter as English tweets, which were later tagged by our assistants as indigenous.

Figure 3 gives an indication of the amount of foreign and discarded content, based on the number of words tagged within either <quote type="foreign website">, <foreign type="poster origin">, or <quote type="potentially spam"> in the first iteration of corpus annotation, or the website links that were found to be foreign.[7] The highest percentage of discarded content is in three corpus categories, i.e. comments, web forums, and websites. In terms of countries, the highest amount of discarded content occurs in Bangladesh, with forums and tweets having the highest percentage overall in their respective category, and websites having the second highest position. In the case of the Pakistani comments, most of the removed content was posted by Indian commenters, and the inverse was true for the Indian comments. In the case of the Bangladeshi comments, there were many spam or advertisement related comments as well as foreign commenters. The main reason of copied content in forum posts was news articles and other non-local content that was copied from other websites and posted as forum posts. Such forum posts were generally several hundred words long. Bangladeshi tweets generally contained content that was either advertisement or spam tweets, e.g. crypto currency related repetitive tweets, motivational and inspirational quotations, or otherwise repetitive tweets with little variation in content posted by freelance Bangladeshi workers in English.[8] About one third of Sri Lankan website links were marked as foreign by

5 The same procedure was also followed for the other three tags as shown in Table 4, last four rows.

6 In addition to the verification of website origin as described in the previous section, websites and blogs were checked a second time and 'repaired' at the annotation stage just like other corpus categories to exclude foreign/copied content, non-local links, and spam web pages.

7 The purpose is to provide an estimate of how much foreign or discarded content can be expected, so we do not include the absolute number of words or percentages that we had to discard during the many iterations of corpus annotation and file repairing for each category and country.

8 For example, the following rather generic type of tweet could be repeated several times: "…Fiverr promoted me to a Level One seller! Proud to be part of a platform that recognizes hard work. check out my Gigs…"

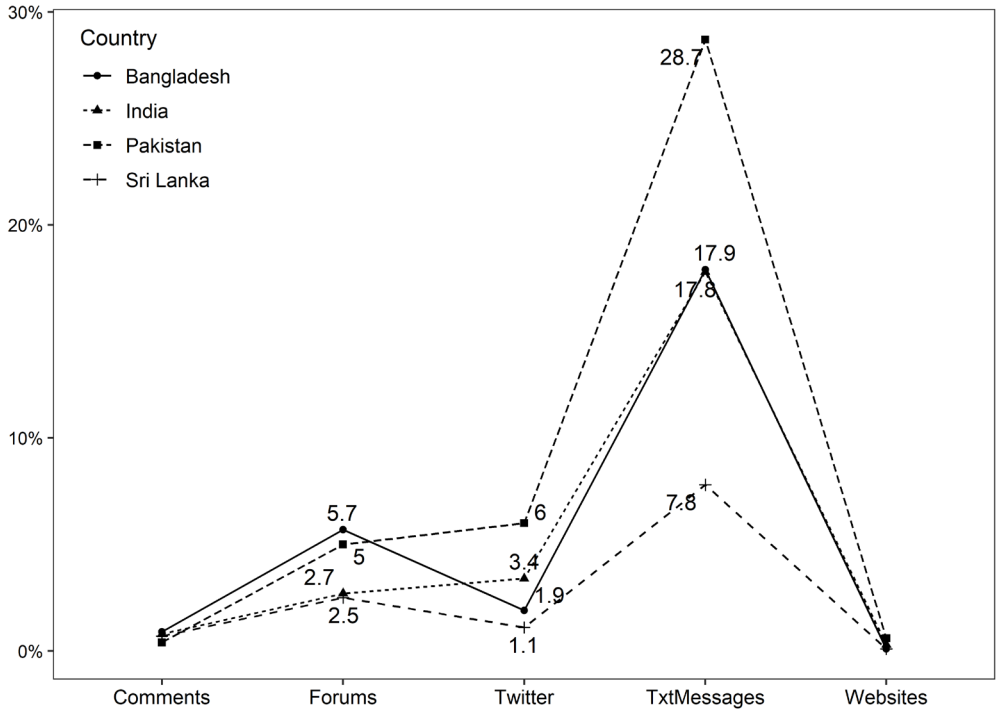


Figure 2. Percentage of indigenous words in various categories of the corpus in the four South Asian countries.

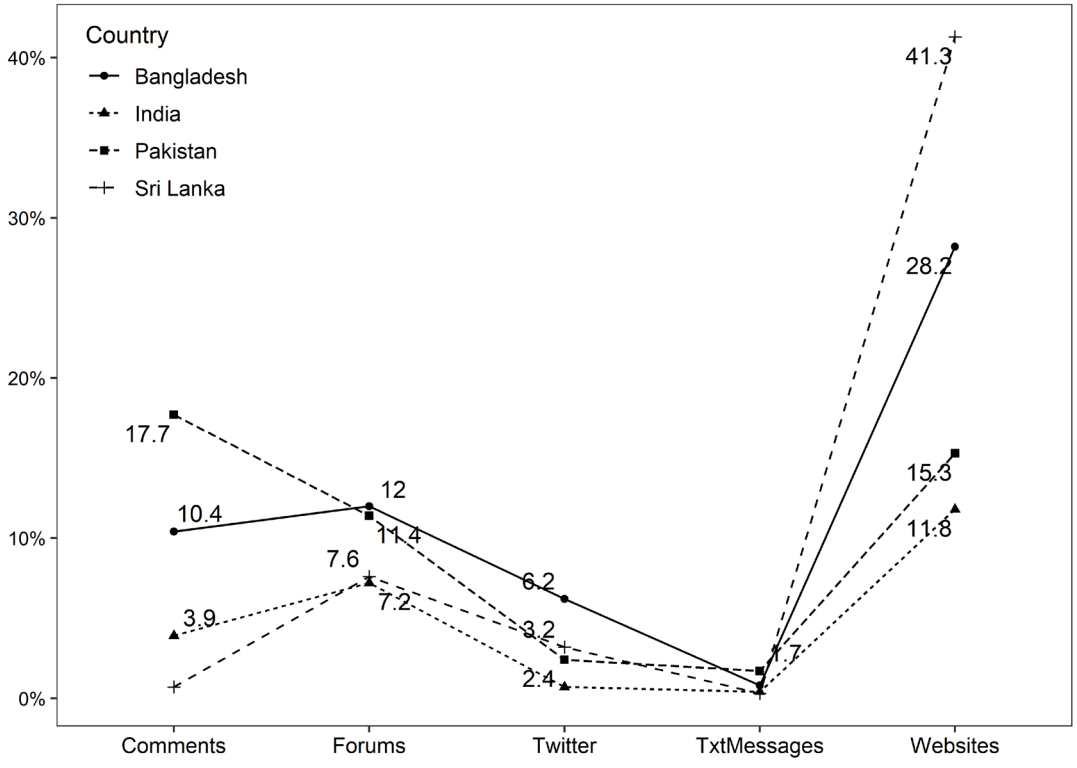


Figure 3. Estimated percentage of foreign and discarded content (number of words) in different corpus categories in the four South Asian countries.

our freelance assistants. An additional 10% or so (in terms of word count) comes from the next phase when content copied from foreign websites was tagged. The main reason of foreign content was links from missionary websites that were flagged as foreign by our Sri Lankan assistants. Many foreign pages in Bangladesh were country-specific pages of international websites with the same content also present in other countries, e.g. the Indian or Sri Lankan page of the same website, and news reports that were adopted and published from international sources.

3.5 The use of freelance assistants

Since we have used freelance helpers from the four South Asian countries extensively in our data collection and annotation process, it seems fair to comment on our experience on the usage of such resources in linguistic projects. Overall, we consider the use of freelance assistants as a mixed experience. They are an attractive alternative to student assistants that are expensive and whose hiring requires a certain bureaucratic procedure. In the case of annotation (e.g. annotating indigenous content), it is advantageous to have the help of freelance assistants as knowing the local language is necessary and most of the times finding a student assistant with such linguistic competence in the said language is difficult. Another advantage is that the task can be performed quite quickly by freelance assistants as compared to a linguistics student who is working part-time, i.e. a few hours a week. Additionally, freelance assistants - when they have some linguistics training can be a very good resource for long term cooperation. For example, we found such freelance assistants from Sri Lanka and Bangladesh who have helped us as informants in additional questions that we had about the data (cf. footnote 13 about *ne* in the pilot study).

On the negative side, there can be issues like the quality of work. For example, in our case, a freelance assistant from Bangladesh submitted Facebook Messenger conversations that included people from other countries as well, which was against the instructions we gave them. We came to know about this quite late in the data collection phase, and had to replace those files. We could not recover the amount paid to the said person because the time duration in which a dispute could be launched at upwork.com had expired. Additionally, freelance assistants will generally be non-experts in terms of linguistic research methods. The researcher should be able to clearly explain to the non-expert about the linguistic task in a very non-technical and layman terminology. Otherwise, there may be unforeseeable consequences (cf. the next subsection on <indig> vs <foreign> tags). Lastly, freelancers are generally paid on a per-task basis, as opposed to per-hour. It is upon the freelancer how much time they spend on the task. It may as well be the case that the said assistant does not spend enough time on the task and skips some things, e.g. a word in an indigenous language is not tagged in <indig></indig>, hence compromising the quality of work.

In conclusion, we believe that the use of freelance assistants should be welcomed with caution. They can be an invaluable resource in very tedious tasks in linguistic projects as well as enormously helpful as local informants. At the same time, the researcher needs to be aware of the potential pitfalls in terms of the quality of work, potential deception etc.

3.6 Interim summary

The SAOnE corpus is a multi-million words corpus of South Asian online communication in English with certain advantages over currently available datasets. The extensive effort put in verifying the sources and removing copied content makes it relatively more reliable, in that it gives more confidence that the data actually originate from a given country, e.g. Bangladesh, and not a neighbouring country like India or another foreign country. The presence and tagging of indigenous content allows the study of phenomena like code-switching. The estimates about the presence of foreign (i.e. non-local) content made based on this dataset, especially in blogs and websites, could be useful for future researchers in evaluating the results from bigger corpora like GloWbE. On completion, this corpus (except the tweets) will be made available through an online portal with the aim of providing another valuable resource to the research community. Apart from the advantages, like any existing resource, this dataset also has its limitations. The process of annotation was performed by assistants who were mostly not trained in

linguistics, which could affect the quality of annotations. The annotation of indigenous and checking for foreign/copied content is by no means perfect or complete, because of the size of the corpus and the limited resources that we had. There are bound to be tagging inconsistencies; for example, Hindi phrases in Sri Lanka were tagged as foreign by local freelance assistants but our Bangladeshi freelance assistants tagged Hindi/Urdu phrases or words as indigenous in Bangladeshi tweets and web forums etc., mainly due to more familiarity with Hindi and due the influence of Bollywood. The lack of text messages from the USA and UK is also a shortcoming of this dataset. Finally, any additional metadata, apart from URLs, and annotations like CMC spelling corrections could make this dataset more useful.

4 Pilot study

In this section we present the pilot study of the three discourse particles *yaar*, *na*, and *neh* from the four South Asian countries included in SAOnE. The following sections provide the background and the results of the study.

4.1 Discourse particles

Discourse markers are linguistic items that exist outside of the core syntax of the clause, and are optional in terms of meaning, i.e. they do not create meaning but explicate the already available connections in the text (Brinton 1996: 267). They perform functions like "connectivity, turn taking, and stance or propositional attitude" (Ler Soon Lay 2006: 153). Discourse markers can be multiword as well as single-word linguistic items. Many single-word or monomorphemic discourse markers are referred to as discourse particles. They are a subset of discourse markers and perform various functions of the former (Ler Soon Lay 2006: 153). Discourse particles can occur utterance finally, as well as in the middle, or at the beginning of the utterance, as our data will show below.

Discourse markers have attracted the attention of linguists, especially in pragmatics, over the last few decades, as noted in Lange (2009). They have also become a topic of interest in the context of language contact. Matras (2000: 517) theorises that due to the discourse organisational function of discourse markers, it is more attractive for language users to borrow these expressions from one language to another to reduce the cognitive load of thinking and speaking along with monitoring and directing of discourse. Matras (2009: 157) shows that discourse markers are number three in the hierarchy of most borrowable items after nouns and conjunctions, and verbs. The relationship between multilingual discourse markers, and hence discourse particles, and borrowing makes them relevant to the study of varieties of English around the world, because many of these non-native varieties exist in a language contact situation with indigenous languages.

Indeed, the study of discourse particles has increasingly awoken the interest of many researchers of World Englishes over the last two or so decades. Singapore English is one of the most researched varieties of English in terms of discourse particles (e.g. Ler Soon Lay 2006). However, there have been studies on other varieties of English too, for example Indian English. Lange (2009) analyses the discourse particle *yaar* in the spoken component of ICE India. *Yaar* is a word that is native to Hindi/Urdu as well as to other South Asian languages like Punjabi. *Yaar*, which literally means *dude* or *man*, also contains sexual connotations in its semantics. *Yaar* is a friend as well as a lover (Afzal 2014: 70). Punjabi elders prohibited their daughters the use of this word, because prostitutes have *yaar*s and not pious women (Shakir in print). Lange's (2009) data shows that in about half of the occurrences *yaar* is used as a discourse particle to emphasise meaning and to express the speaker's attitudes. Similar observations have been made by Shakir (in print) in Pakistani English CMC.

Another discourse particle native to Hindi/Urdu and also used in Indian English is *na*. The Hindi question tag, which is used with positive or negative statements, was first reported by Valentine (1991: 331) in connection with Indian English. Lange (2009) confirms this finding in her ICE data. The same discourse particle, which is borrowed from Urdu, is used in Pakistani English tweets and blog comments (Shakir in print). In the context of South Asia, it is also noteworthy to mention that *na* is also

a Bangla word which literally means *no* (Dastidar and Mukhopadhyay 2013). In addition to its function as a negation marker, it is also used as a tag question, for example "ar ekTuo chocolate nei, *na*? (there is no more chocolate, *right*?)" (Dastidar and Mukhopadhyar 2013: 817). *Na*, thus, can potentially be used by Indians, Pakistanis, as well as by Bangladeshi users of English in spoken communication as well as in CMC. The negation marker *neh* from Sinhala (the majority language spoken in Sri Lanka) is also of interest due to similar properties like *na*. *Neh* /næhæ/, which means 'not' in Sinhala (Hilpert 2006: 50), as well as the forms /ne:/ and /ne:də/ ('no') can be used to create interrogatives from declarative sentences.[9] Ranaweera (2022) lists *neh* as a nativized discourse marker in Sri Lankan Youtube interviews of 12 men and 12 women. The study of *neh* in a larger dataset of CMC registers will certainly shed more light on the usage of this indigenous item in informal Sri Lankan English communication.

The above discussion demonstrates that South Asian users of English can potentially use discourse particles borrowed from indigenous languages, and it need not be limited to India and Pakistan. However, the informal contexts in which such expressions are more likely to occur, i.e. face-to-face or spoken conversations, do not exist in all South Asian countries in English. Even if they do, the general unavailability of such data makes it difficult to study these features. However, CMC registers like text messages, Twitter, web forums, and comments are interactive to a certain extent just like spoken discourse. The same features can be studied through CMC originating from the above-mentioned South Asian countries. To the best of our knowledge no such study exists as yet, so the present pilot study intends to fulfil this research gap.

As per this discussion, our aim in this section is to analyse the use of three discourse particles *na*, *neh*, and *yaar* in the SAOnE corpus with a view to find out any differences between the four South Asian varieties of English included in this corpus. However, before that we present the usage patterns of these particles in the South Asian component of the GloWbE corpus to highlight why a corpus like SAOnE with a wider range of internet-based registers is more suitable to study such phenomena.

4.2 Discourse particles in GloWbE

GloWbE includes all four South Asian countries that are also included in the SAOnE corpus. The total number of words included in the South Asian subsection (Bangladesh, India, Pakistan, and Sri Lanka) of the corpus is approximately 234,039,410. There are 2 text categories in the corpus: Blogs and General websites. Table 5 exhibits the overall occurrences of each of the three discourse particles in the South Asian section. As can be observed, the particle *yaar* has the highest number of true positives. It is present in all 4 countries: BD=2, IN=64, PK=36, and LK=4. The other notable aspect is that about half of the concordance lines we analysed for the particle *na* were the English particle *na* as in gon*na* and wan*na*. We only found 5 occurrences (BD=1, IN=2, PK=2,) of this particle in the remaining 500 or so concordance lines. *Ne* only has 1 occurrence, which is unsurprisingly from the Sri Lankan part of GloWbE.[10]

As *yaar* is the most frequently occurring discourse particle in our analysis, we further investigate the category-wise distribution of it as shown in Table 6. The majority of occurrences of the particle are in the general websites section. On a closer look, we find that apart from blogs and general websites, the corpus also contains forum websites, which are mostly listed as websites but at times also as blogs in the corpus. For example, we found 7 occurrences of *yaar* on domains that include the word 'forum'. These occurrences are on two forum websites categorised as Indian, i.e. india-forums.com and r2iclubforums.com, which are still accessible. The highest number of *yaar* occurrences are on India-forums.com (6): the forum is categorised 5 times under general websites and once as a blog. We came across another forum desitvforum.net, where the particle is used in a multiword expression, i.e. '*Arrey yaar* (oh dude!), there was a…'.[11] This third forum, categorised as a Pakistani blog, does not exist anymore. A look at

9 Thanks to our Sri Lankan informant Shiny Adjumian for this information.

10 Spelling variant *neh* has 8 occurrences in GloWbE (South Asia) and none of them has been used as a discourse particle.

11 Multiword expressions involving discourse particles have not been included in frequency counts shown in Tables 5 and 6. The same is true for our analysis of these particles in SAOnE.

Table 5. Frequency and other statistics of the three discourse particles in GloWbE.

Discourse particle	Total concordance lines	Randomly analysed concordance lines	True positives	Words per million*
na	13103	1000	5	0.0214
ne	2064	200	1	0.0043
yaar	394	200	106	0.4529

*Based on total words in the 4 countries 234,039,410

Table 6. Category-wise distribution of *yaar* in GloWbE.

Country	General websites	Blogs
BD	1	1
IN	44	20
PK	27	9
LK	2	2

its Wayback Machine instance from 2010 shows that it is most probably an Indian forum with a mix of Pakistani members and a sub-forum titled 'Pakistani Forums'.[12]

Our analysis of these discourse particles in the GloWbE corpus brings forth the following findings. Firstly, the particles *na* and *ne* are not very common in the GloWbE data, which is partly due to the under-representation of interactive registers. Additionally, these two particles are hard to filter out due to their identical form with other English words, i.e. *na* with gon*na* and *ne* with NE as an abbreviation of North East. Despite low frequency, *yaar* is easy to capture probably because it has no homonym in English. Secondly, there are not just blogs and other websites, there are also web forums buried in either of the two categories, as also found out by Biber et al. (2015) for the five native varieties of English. These web forums can either be categorised under Blogs or General websites, with no clear-cut criteria deducible. Thirdly, the web documents categorised under one country may as well be from another country (cf. above). Due to highly mixed text categories, under-representation and no categorisation of even openly accessible interactive registers like web forums, and at times unreliability regarding the country of origin in the GloWbE corpus, SAOnE is a more attractive option to study phenomena like discourse particles. We present the data and results from SAOnE in the following sections.

4.3 Data and methodology

Concordance lines were extracted from the corpus when Twitter data was still being annotated. A custom written Python script was used for this purpose.[13] The search query included case insensitive occurrences of the three discourse particles, i.e. *yaar*, *na*, and *neh*, including spelling variations like repetition/lengthening (*naa* instead of *na*) or reduction (*yr* instead of *yaar*). Additionally, various variations of *neh* like *noh*, *ne* were also included. The English negation marker *no* and its informal version *nah* also occurred as tag questions, mainly in the Sri Lankan data but in other countries, too. However, we did not consider these discourse-level items as variants of *neh* due to their English origin.[14]

12 See https://web.archive.org/web/20100611230514/http://www.desitvforum.net/forum/.

13 The search query was adapted to include particles that were not yet tagged in <indig></indig> tags.

14 *Nah* and *na* are very similar to each other. It is very much possible that some tokens of *nah* could be used in the sense of *na*. However, we use the frequency of occurrence of *nah* to exclude it from our analysis. *Nah* has one appearance in Pakistan and zero occurrences in India. It has a frequency of three in Bangladesh and 86 occurrences in Sri Lanka. Since this variant does not occur in India and negligibly in Pakistan, the two countries where Hindi/Urdu is spoken, we can safely say that *nah* and *na* are different entities. Additionally, *nah* is most probably related to English and, based on its frequency in Sri Lankan data, there are chances that it has also links with *ne* in Sri Lanka, but we chose to be on the cautious side and excluded it completely from our analysis.

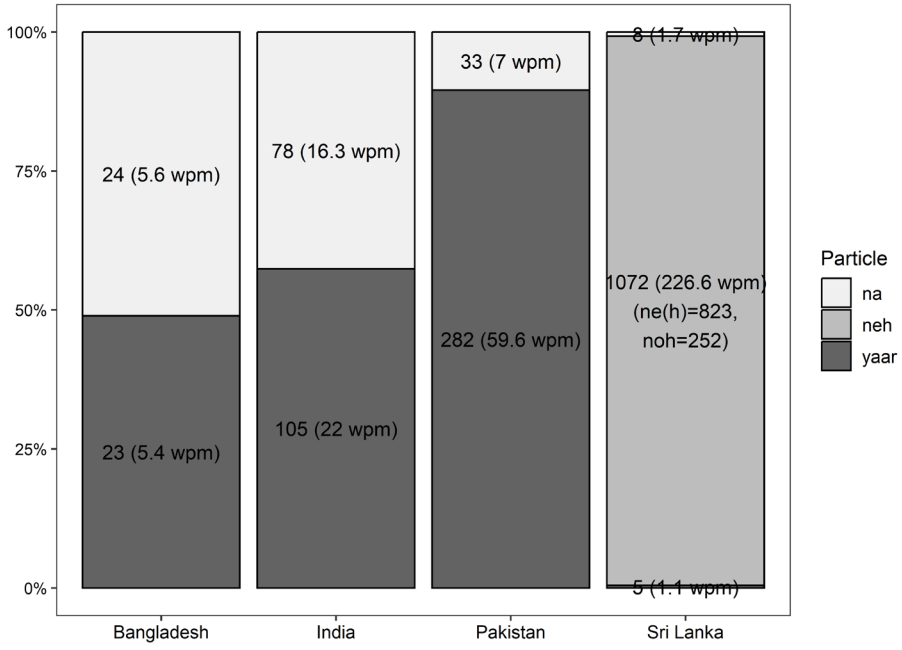


Figure 4. Distribution of discourse particles in the four South Asian countries of the SAOnE corpus (labels consist of raw frequency and words per million according to each country in brackets).

The concordance lines were manually scanned to verify the occurrence of the given discourse particle in English communication. For the discourse particle to be included in our final analysis, it had to occur in a clause or utterance that could be clearly understood as an English clause/utterance. After the manual verification, the concordance lines were tagged for the type of discourse particle (i.e. *na, neh*, or *yaar*) and the position of the discourse particle in the clause or utterance (i.e. start, middle, or end). The data was then tabulated according to country and register category to create bar plots that are presented below.

4.4 Discourse particles in SAOnE

The usage of the three discourse particles in terms of frequency and words per million is shown in Figure 4. As it is clear from the bar plot, the two Hindi/Urdu discourse particles, i.e. *yaar* and *na*, are present in all South Asian countries, though they are prominently used only in Indian and Pakistani data. Even in these two countries, their distribution is specific, i.e. *yaar* is more common in both varieties and *na* is only relatively common in India. Bangladesh has a couple of dozen occurrences of each of the two particles, while Sri Lanka has only a few. The use of *neh* in the Sri Lankan data is much more frequent than any of the other two discourse particles. Examples 1–5 show the use of these discourse particles in India, Pakistan, and Sri Lanka. The three variants of *neh* have been listed in 3–5. As seen especially in 3 and 4, the two versions of *neh* function as question tags irrespective of the spellings.

1. $B: If u will get..then plz send *yr* (=yaar). (WhatsApp, India)
2. But this not actual human 😂 ve to mention a real one *na*? (Twitter, Pakistan)
3. This is not the first instance *noh*? (Comments, Sri Lanka)
4. $L: So our classes will start soon *neh* guys? 😂 (Text messages, Sri Lanka)
5. I guess u can't even look at words normally after getting used to this *ne*. (Twitter, Sri Lanka)

Figure 5 depicts the register-wise distribution of the three discourse particles in the four South Asian countries. In the case of Sri Lanka, India, and Pakistan, there exists a relationship between the words per million and the interactivity of the register category. Hence, the highest frequency of discourse

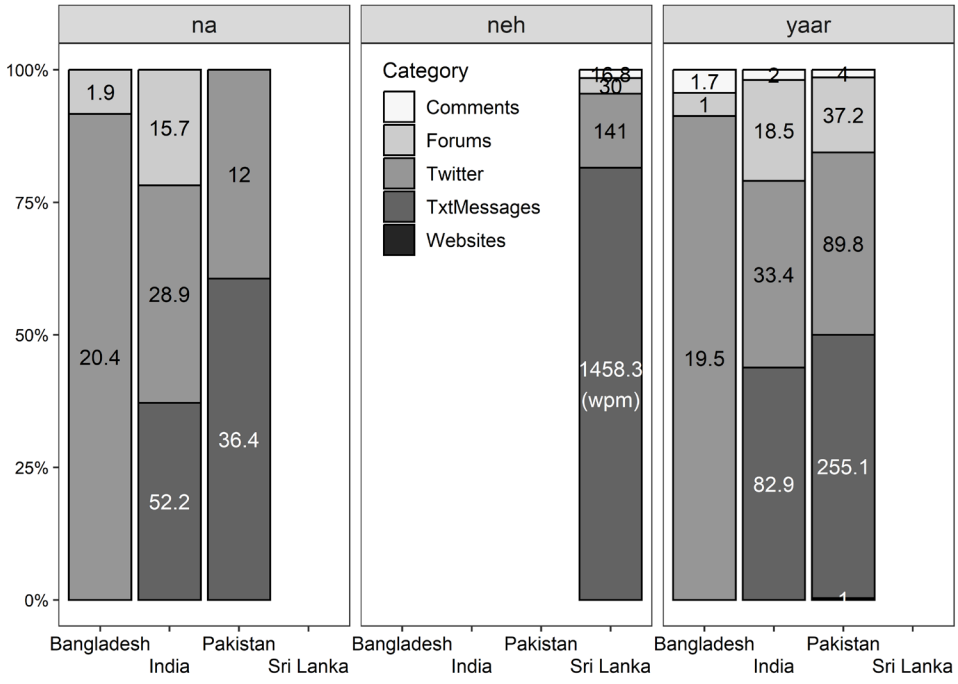


Figure 5. Distribution of discourse particles in the 5 register categories of the SAOnE corpus (words per million according to each register in each country as labels).

particles in each of the three countries occurs in text messages followed by tweets and web forums. In the case of Bangladesh, most occurrences of *yaar* and *na* are in tweets, with no occurrences in text messages. The absence of *na* in Bangladeshi text messages indicates that the Bangla negation marker *na* is not borrowed in the English communication of Bangladeshi users. The only instances where *na* was found in our text messages data were within indigenous Bangla utterances or multiword expressions involving *na*.

The use of the discourse particles *na* and *yaar* in Bangladeshi English communication on Twitter merits a deeper look into the background of these users. We have mentioned above that it was comparatively harder to find enough users tweeting in English than in Bangladesh. A sizeable portion of the users included in our data are fan accounts of foreign and international celebrities, for example South Korean boy bands, Indian celebrities, and in at least one case a Pakistani drama fan account. These users, especially in the case of Indian and Pakistani fan accounts, sometimes use linguistic items borrowed from Hindi/Urdu, which also include the two discourse particles under discussion. This can happen when these fan accounts are interacting with their respective celebrity's Twitter account or possibly when they are communicating with other fan accounts from within or outside of Bangladesh. There are 23 Twitter accounts in our data who have either used *na* or *yaar* at least once in their tweets. Nine accounts are self-proclaimed fan accounts, of which seven belong to fans of Indian celebrities and one is a Pakistani drama fan account.

The following examples show some of the uses of these two discourse particles by Bangladeshi fan accounts and by a Sri Lankan Twitter user. In Examples 6 and 7, the discourse particle *na* is used in an emphatic way. The use of emojis enhances the effect. In Example 8 and 9, the users are referring to relevant celebrity accounts by tagging them. The former is an Indian celebrity, while the latter is a Pakistani actor's account. *Yaar* is being used to refer to the celebrity and to show the fans' affection. The last example, i.e. Example 10, shows a reply to another Twitter account which is most probably a Pakistani user. The use of *yaar* creates a familiarity between the Pakistani drama fan from Bangladesh and the person from Pakistan who also appears to be a fan. The same type of accommodation is visible

in Example 11, where a Sri Lankan user is saying *happy birthday* to an Indian who might be a colleague or friend. The use of *na* appears to be only for engaging the listener/reader and not as a tag question, as Valentine (1991) has also found in her spoken English data from female Indian speakers.

6. Dont sleep *na* 🥺 (Twitter, Bangladesh)
7. @XxxxxxXxxxxXX Drop a selfie *naaa*😭🫳 (Twitter, Bangladesh)
8. @XxxxxXxxxx22 Thank you so much *yaar* 🧡🧡🧡🧡🧡 (Twitter, Bangladesh)
9. @xxxxxxxxxx @xxxxxxxxxx Can't wait *yaar* 😭 This scene 🥺😭 (Twitter, Bangladesh)
10. @XxxxxXxxxx *Yaar* we moved on way earlier. (Twitter, Bangladesh)
11. @Xxxxxx @ Happy Birthday *na*!!!may god bless u,stay happy & healthy😇❤️🏃 (Twitter, Sri Lanka)

These discourse particles can be used at various positions in the clause or utterance. Figure 6 demonstrates the percentage of the usage of each discourse particle in various positions in the four South Asian countries. The discourse particles *na* and *neh* have a very similar distribution in terms of position, i.e. both are used predominantly at the end of a clause or utterance. Example 12 shows the emphatic use of *na*, which is similar to Example 11 above. Example 13 from Pakistani text messages shows its use as a confirmatory tag question. A similar usage for *neh* can be seen in Example 14, where it can be directly translated as *isn't it*. The function of *neh* in the middle of the sentence in Example 15 is not quite clear: it could be for emphasis and interlocutor engagement or it could be a confirmatory tag question, or both.

12. $C1: […] $S3P18 just ask him *na* if he can.
    $O1: Yeah
    $O1: Plz (Text messages, India)
13. $W: $A only verbal *na*? (Text messages, Pakistan)
14. $A: start time is 5:30 *neh*?
    $B: Yeah (Text messages, Sri Lanka)
15. $W: Will have a video chat *neh* altogether (Text messages, Sri Lanka)
16. $D: mam lost her mind
    $S: hahaha true
    $W: *yaar* seriously she said she wanna see total 100 students in her next class [...] (Text messages, Pakistan)
17. […] Beautifully written...u penned down Riddhu inner turmoil so well dear ...the battle bet the her and mind was great
    Plzzz Gur do continue this one *yaar* 🙏🙏🙏 (Forums, India)

As appears from the figure, *yaar* is much more flexible as compared to the other two discourse particles and is more prominently used at the beginning of the clause or utterance in the Pakistani data as compared to the Indian data. *Yaar* at the beginning and at the end of utterance has been shown in Examples 16 and 17. In the Pakistani text message exchange in 16, three female students are talking about their teacher. It is not quite clear who is being referred to with the word *yaar* here, so it could be both functions: the primary vocative function and the bleached discourse particle function, as Lange (2009: 2017) has noted. In 17 an Indian forum post shows the use of *yaar* at the end. It could be argued that the person being referred to as *Gur* is also being called *yaar*. At the same time, the emphatic function cannot be ruled out here, either. Finally, irrespective of the position in clause or utterance, we have noted a number of positive, neutral, and negative adjectives occurring before *yaar* where the discourse particle is clearly used to emphasise meaning. Some adjectives and expressions are as follows: *good, nice, literally, wow, same, seriously, great, sad, sexy, shit, amazing, awesome, bad, better, mind-blowing, cool, exactly, fuck, wonderful* etc. We argue that in most cases it is probable that both functions of *yaar*, i.e. the primary vocative and discourse level function of emphasising meaning, are present.

4.5 Discussion and implications

We presented two analyses in the previous subsections. The analysis based on GloWbE data highlighted that SAOnE is a more appropriate choice for such a study due to the presence of informal/interactive internet-based registers, among other reasons. The usage patterns of three discourse particles in the SAOnE corpus have shown that the discourse particles occur in text messages, Twitter, web forums,

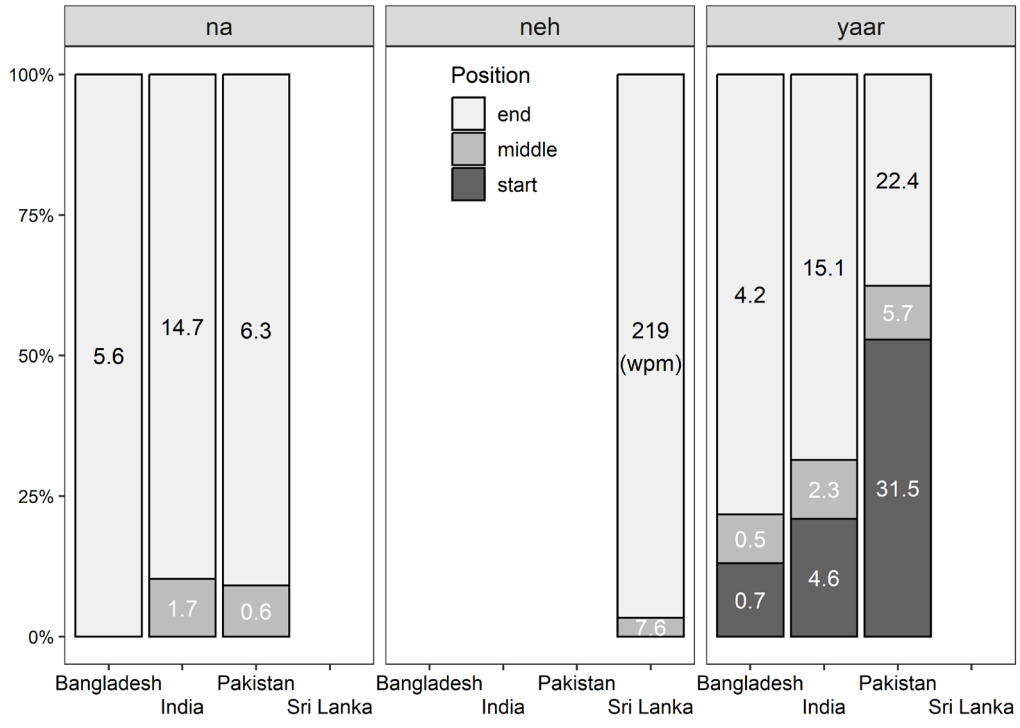


Figure 6. Distribution of discourse particles according to their position in clause/utterance in the SAOnE corpus (words per million according to each country as labels).

comments, and general websites in descending order of frequency. The two Hindi/Urdu discourse particles, i.e. *na* and *yaar*, occur not only in India and Pakistan but also in Bangladesh and Sri Lanka. Generally, it can be assumed that in such cases the tweet is somehow related to India or Pakistan. India and Pakistan have different frequencies of the two particles, i.e. *na* is relatively more common in the Indian data and *yaar* in both the Indian and the Pakistani data with slight differences. The Sinhala-based discourse particle *neh* is exclusive to Sri Lanka. *Na* and *neh* occur most prominently at the end of a clause or utterance. *Yaar* is much more flexible and more commonly occurs at the clause or utterance beginning in the Pakistani data. Overall, it can be said that *yaar* and *na* involve an emphasising function, especially *yaar*, even when it is being used as a vocative. *Neh*, while being restricted to the Sri Lankan data, is the most frequent of all, which indicates that it is a prominent feature of Sri Lankan English communication.

The results of our analysis show some regional trends as well as some country-specific traits of the CMC communication of the four South Asian countries, i.e. Bangladesh, India, Pakistan, and Sri Lanka. At the regional level, we see similarities in the usage of discourse particles and the influence of bigger South Asian countries, especially India, on other South Asian countries, namely Bangladesh and Sri Lanka. The use of discourse particles by Bangladeshi fan accounts shows how Indian culture through its film industry influences other neighbouring countries. The same can be said for Pakistani dramas, though to a much lesser extent as compared to India. Additionally, the same kind of linguistic accommodation is visible in the very rare use of Hindi/Urdu discourse particles in Sri Lanka, which again indicates the cultural influence of bigger South Asian countries like India. The common practices in the use of Hindi/Urdu indigenous discourse particles in South Asia depict that multilingual resources are also transferred in the language contact between different South Asian varieties of English. Hence, to understand the relationship of various South Asian Englishes and to study epicentre roles in South Asia (e.g. Hoffmann et al. 2011), it would also be beneficial to look at the role of multilingual resources like indigenous discourse particles.

The differences in the use of discourse particles in these data show the unique linguistic practices of the four South Asian countries, which in at least this case are also partly influenced by the substrate languages used in these countries. In the case of India and Pakistan, the Hindi/Urdu discourse particles show different distributions. These two varieties (i.e. Hindi and Urdu) are very similar linguistically, but the speech communities in which they are used (i.e. India and Pakistan respectively) have certainly diverged from each other over the past 70 or so years. More research is needed to further explain these differences, including research on these discourse particles in their native environment, i.e. in Hindi and Urdu. The use of the Sinhala discourse particle *neh* in Sri Lanka shows the influence of the local majority language on the Sri Lankan variety of English, even though there are similarities between *neh* and the Hindi/Urdu discourse particle and negation marker *na*. Lastly, although the same negation marker is available in Bangla and also reported to be used in similar functions as *na* in Hindi/Urdu (Dastidar and Mukhopadhyay 2013), at least these data show that Bangladeshi users of English do not employ the discourse particle in their CMC interactions. It could be partly attributable to the type of data we analysed for Bangladesh: even the most informal register, i.e. text messages, has conversations that are either up to 90 percent in Bangla or when the majority of the communication is in English, the topics are quite formal, e.g. study, so no code-switching occurs where a discourse particle like *na* could be found. Looking at it in another way, this shows that English is not nativised in Bangladesh in the way it has been in other South Asian countries like India.

The pilot study presented in the previous subsection also has implications for the type of data that we have analysed, i.e. computer mediated communication. These data have shown that text messages are a very informal and interactive register. Though they cannot by any means replace spoken conversations, the register can still be a very good source of data when spoken conversations are not available and the researcher wants to study linguistic features that only occur in informal situations and registers. Twitter can be a good source to study the influences and interactions of various South Asian and other English-speaking countries with each other. The other registers included in the corpus, i.e. web forums, comments, and general websites including blogs, can also show regional and country-specific trends, which include multilingual resources as well as innovations in the local variety of English. The comparison of the GloWbE results with those from SAOnE has further emphasised the usefulness of a more diverse corpus of CMC registers. In summary, we hope to have shown the importance of compiling such corpora for South Asia as well as other countries where English as a non-native language has an important role in day-to-day communication to understand the influence of the internet and the spread of and/or variation in English.

5 Conclusion and future research directions

We have presented a new corpus of English CMC from four South Asian countries, i.e. Bangladesh, India, Pakistan, and Sri Lanka. The dataset has been collected using a semiautomatic approach and we have tried to eliminate non-local and copied content to make it as reliable as possible. The corpus also has its weaknesses, for example the lack of text messages from the countries included for comparison, i.e. the United Kingdom and the United States of America, which would have been useful for our planned analyses as well as for other researchers after its release. To highlight the importance of such a dataset, we have conducted a pilot study which has provided new and unique perspectives on the use of multilingual resources in and among the four South Asian countries. In terms of future research directions, we aim to study other aspects of language variation, for example grammatical variation in comparison with the USA and UK data included in the corpus for this purpose and a deeper look at the code-switching practices of the South Asian users in the context of CMC. We also aim to finalise the corpus and make it available to the World Englishes community so that this resource can be used for further research.[15]

15 The public release of the corpus is planned for after completion of our own scheduled analyses. In this process, we also hope to further rectify any issues that might have been overlooked in the data collection and annotation phases.

Acknowledgements

We would like to acknowledge that this project has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 452561886. We would also like to extend our gratitude to the student and freelance assistants who helped in this project and the anonymous reviewer whose comments were very helpful.

References

Afzal, Ahmed. 2014. "Being gay has been a curse for me": Gay Muslim Americans, narrative and negotiations of belonging in the Muslim *ummah*. *Journal of Language and Sexuality* 3 (1): 60–86. https://doi.org/10.1075/jls.3.1.04afz.

Biber, Douglas and Jesse Egbert. 2018. *Register variation online*. 1st edn. Cambridge: Cambridge University Press.

Biber, Douglas, Jesse Egbert and Mark Davies. 2015. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora* 10 (1): 11–45. https://doi.org/10.3366/cor.2015.0065.

Bohmann, Axel. 2019. *Variation in English worldwide: Registers and global varieties.* 1st edn. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108751339.

Brinton, Laurel J. 1996. *Pragmatic markers in English: Grammaticalization and discourse functions*. Berlin: De Gruyter. https://doi.org/10.1515/9783110907582.

Chowdhury, Raqib and Ariful Haq Kabir. 2014. Language wars: English education policy and practice in Bangladesh. *Multilingual Education* 4 (1): 1–16.

Danilak, Michal Mimino. 2021. *langdetect*. Python. https://pypi.org/project/langdetect/.

Dastidar, Rimi Ghosh and Sibansu Mukhopadhyay. 2013. Utterance discourse and meaning: A pragmatic journey with the Bangla discourse particle /na/. In R. Prasath and T. Kathirvalavakumar (eds.). *Mining intelligence and knowledge exploration* (Lecture Notes in Computer Science). Vol. 8284, 814–822. Cham: Springer. https://doi.org/10.1007/978-3-319-03844-5_80.

Davies, Mark and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide* 36 (1): 1–28.

Hilpert, Martin. 2006. The lexical category auxiliary in Sinhala. In R. Englebretson and C. Genetti (eds.). *Proceedings from the Workshop on Sinhala Linguistics*. Vol. 17, 49–67. Santa Barbara: Department of Linguistics, University of California.

Hoffmann, Sebastian, Marianne Hundt and Joybrato Mukherjee. 2011. Indian English – an emerging epicentre? A pilot study on light verbs in web-derived corpora of South Asian Englishes. *Anglia – Zeitschrift für englische Philologie* 129 (3–4): 258–280. https://doi.org/10.1515/angl.2011.083.

Kirk, John and Gerald Nelson. 2018. The International Corpus of English project: A progress report. *World Englishes* 37 (4): 697–716. https://doi.org/10.1111/weng.12350.

Kouzis-Loukas, Dimitrios. 2016. *Learning Scrapy*. Packt Publishing Ltd.

Lange, Claudia. 2009. 'Where's the party *yaar*!' Discourse particles in Indian English. In T. Hoffmann and L. Siebers (eds.). *World Englishes – problems, properties and prospects. Selected papers from the 13th IAWE conference*, 207–22. Amsterdam: Benjamins.

Leitner, Gerhard. 1991. The Kolhapur Corpus of Indian English – intravarietal description and/or intervarietal comparison. In S. Johansson and A. Stenström (eds.). *English computer corpora: Selected papers and research guide*, 215–232. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110865967.215.

Ler Soon Lay, Vivien. 2006. A relevance-theoretic approach to discourse particles in Singapore English. In K. Fischer (ed.). *Approaches to discourse particles*, 149–166. Amsterdam: Elsevier. https://doi.org/10.1163/9780080461588_010.

Loureiro-Porto, Lucía. 2017. ICE vs GloWbE: Big data and corpus compilation. *World Englishes* 36 (3): 448–470. https://doi.org/10.1111/weng.12281.

Mair, Christian. 2013. The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34 (3): 253–278. https://doi.org/10.1075/eww.34.3.01mai.

Mair, Christian. 2015. Response to Davies and Fuchs. *English World-Wide* 36 (1): 29–33.

Matras, Yaron. 2000. Fusion and the cognitive basis for bilingual discourse markers. *International Journal of Bilingualism* 4 (4): 505–528. https://doi.org/10.1177/13670069000040040701.

Matras, Yaron. 2009. *Language contact*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809873.

Mukherjee, Joybrato and Marco Schilk. 2012. Exploring variation and change in new Englishes looking into the International Corpus of English (ICE) and beyond. In T. Nevalainen and E. C. Traugott (eds.). *The Oxford handbook of the history of English*, 189–199. Oxford: Oxford University Press.

Mukherjee, Joybrato, Marco Schilk and Tobias Bernaisch. 2010. Compiling the Sri Lankan component of ICE: Principles, problems, prospects. *ICAME Journal* 34: 64–77.

Nelson, Gerald. 2015. Response to Davies and Fuchs. *English World-Wide* 36 (1): 38–40.

Python Core Team. 2019. *Python: A dynamic, open source programming language*. Python Software Foundation. https://www.python.org/.

Ranaweera, Mahishi. 2022. A gender-based analysis of pragmatic markers in Sri Lankan English. In *ICAME 43, Book of abstracts*. Anglia Ruskin University, Cambridge, UK. https://icame.info/wp-content/uploads/ICAME43_Abstracts_v2.pdf.

Sardinha, Tony Berber. 2022. A text typology of social media. *Register Studies* 4 (2): 138–170. https://doi.org/10.1075/rs.22008.ber.

Shakir, Muhammad. In print. Functions of code-switching in online registers of Pakistani English. In G. Wilson and M. Westphal (eds.). *New Englishes, new methods*. Amsterdam: Benjamins.

Shastri, S. V. and Gerhard Leitner. 2002. *The ICE-India Corpus. Version 1*. http://ice-corpora.net/ice/download.htm.

Sridhar, S. N. 2020. Indian English. In K. Bolton, W. Botha and A. Kirkpatrick (eds.). *The handbook of Asian Englishes*, 241–277. Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118791882.ch10.

Twitter. 2022. *Developer terms*. Twitter. https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases. (12 December, 2022).

Valentine, Tamara M. 1991. Getting the message across: Discourse markers in Indian English. *World Englishes* 10 (3): 325–334. https://doi.org/10.1111/j.1467-971X.1991.tb00167.x.